Supplementary Material

Ahrim Youn[*] Richard Simon[*†]

Determination of c_k, d_k, e_k, f_k

The constants c_k, d_k, e_k, f_k can be determined from the gene sequence and genetic code. However, the values of these constants and random variables X_{jk} are ambiguously defined in cases where genes have several alternative transcripts, and where some base pairs belong to different codons in alternative transcripts. For example, some exons have different reading frames in different transcripts of the same gene, thus a base pair can be placed in a different position within a triplet codon in alternative transcripts. Also, when an exon ends within a codon, that codon may change when the exon is spliced with an alternative exon starting with different base pairs. When a codon of the base pair changes, the result of the mutation on that base pair can become different, that is, the same mutation can be nonsilent in one transcript and silent in another. Therefore, the value of the random variables,X_{jk} and the constants c_k, d_k, e_k, f_k may differ depending on which transcript the base pair is assigned to.

In such cases, we follow the annotation rules of the corresponding dataset. For example, Ding *et al.* (2008) considered a mutation as nonsilent if it is nonsilent in any transcript. Therefore, when we apply our method to their dataset, if the value of X_{jk} can take both sts and nts or stv and ntv depending on the transcript, we determine its value to be nts or ntv. Accordingly, we assign to e_k and f_k the maximum of the set of values that they can take. On the contrary, we assign to c_k and d_k the minimum of the set of values that they can take.

[*] Biometric Research Branch, National Cancer Institute, 9000 Rockville Pike, MSC 7434, Bethesda MD 20892-7434, USA

[†] Correspondence should be addressed to R.S. (rsimon@mail.nih.gov)

Derivation of the method of moments estimates $\hat{r}, \hat{p}_m$

Here, we describe the process of deriving the method of moments estimates $\hat{r}$ and $\hat{p}_m$.
For $m = 1,3,5$

$$E(\sum_j \sum_{t_k=m} I(X_{jk} = \text{sts})) = p_m \sum_j q_j \sum_{\substack{t_k=m \\ k\in K}} c_k = p_m \sum_j q_j C_m, \tag{1}$$

$$E(\sum_j \sum_{t_k=m} I(X_{jk} = \text{nts})) = r p_m \sum_j q_j \sum_{\substack{t_k=m \\ k\in L}} e_k = r p_m \sum_j q_j E_m, \tag{2}$$

For $m = 2,4,6$

$$E(\sum_j \sum_{v_k=m} I(X_{jk} = \text{stv})) = p_m \sum_j q_j \sum_{\substack{v_k=m \\ k\in K}} d_k = p_m \sum_j q_j D_m, \tag{3}$$

$$E(\sum_j \sum_{v_k=m} I(X_{jk} = \text{ntv})) = r p_m \sum_j q_j \sum_{\substack{v_k=m \\ k\in L}} f_k = r p_m \sum_j q_j F_m, \tag{4}$$

and

$$E(\sum_j \sum_k I(X_{jk} = \text{iid})) = r p_7 \sum_j q_j |L|, \tag{5}$$

$$E(\sum_j \sum_k I(X_{jk} = \text{fid})) = r p_8 \sum_j q_j |L|, \tag{6}$$

By dividing the equation(2) with (1) and (4) with (3), we obtain the following equations.

$$r\frac{E_m}{C_m} = \frac{E(\sum_j \sum_{t_k=m} I(X_{jk} = \text{nts}))}{E(\sum_j \sum_{t_k=m} I(X_{jk} = \text{sts}))} \qquad \text{for } m = 1,3,5$$

$$r\frac{F_m}{D_m} = \frac{E(\sum_j \sum_{v_k=m} I(X_{jk} = \text{ntv}))}{E(\sum_j \sum_{v_k=m} I(X_{jk} = \text{stv}))} \qquad \text{for } m = 2,4,6$$

Therefore, a natural method of moments estimate for r is

$$\hat{r} = \frac{1}{6}\Big(\sum_{m=1,3,5} \frac{C_m \sum_j \sum_{t_k=m} I(X_{jk} = \text{nts})}{E_m \sum_j \sum_{t_k=m} I(X_{jk} = \text{sts})} + \sum_{m=2,4,6} \frac{D_m \sum_j \sum_{v_k=m} I(X_{jk} = \text{ntv})}{F_m \sum_j \sum_{v_k=m} I(X_{jk} = \text{stv})}\Big)$$

If we add the equation(1) to (2) and (3) to (4), we obtain the following equations.

$$E(\sum_j(\sum_{t_k=m} I(X_{jk} = \text{sts or nts}))) = p_m \sum_j q_j(C_m + rE_m) \qquad \text{for } m = 1,3,5$$

$$E(\sum_j(\sum_{v_k=m} I(X_{jk} = \text{stv or ntv}))) = p_m \sum_j q_j(D_m + rF_m) \qquad \text{for } m = 2,4,6$$

Since $p_1 = 1$,

$$\sum_j q_j = \frac{E(\sum_j(\sum_{t_k=1} I(X_{jk} = \text{sts or nts})))}{C_1 + rE_1}$$

Therefore, a natural method of moments estimate for p_m is

$$\hat{p}_m = \frac{(C_1 + \hat{r}E_1)\sum_j(\sum_{t_k=m} I(X_{jk} = \text{sts or nts}))}{(C_m + \hat{r}E_m)\sum_j(\sum_{t_k=1} I(X_{jk} = \text{sts or nts}))} \qquad \text{for } m = 1,3,5$$

$$\hat{p}_m = \frac{(C_1 + \hat{r}E_1)\sum_j(\sum_{v_k=m} I(X_{jk} = \text{stv or ntv}))}{(D_m + \hat{r}F_m)\sum_j(\sum_{t_k=1} I(X_{jk} = \text{sts or nts}))} \qquad \text{for } m = 2,4,6$$

$$\hat{p}_7 = \frac{(C_1 + \hat{r}E_1)\sum_j \sum_k I(X_{jk} = \text{iid})}{\hat{r}|L|\sum_j(\sum_{t_k=1} I(X_{jk} = \text{sts or nts}))}$$

$$\hat{p}_8 = \frac{(C_1 + \hat{r}E_1)\sum_j \sum_k I(X_{jk} = \text{fid})}{\hat{r}|L|\sum_j(\sum_{t_k=1} I(X_{jk} = \text{sts or nts}))}$$

Derivation of $F_{ij}(x)$

Let W_{ij} be the number of nonsilent mutations that occurred in gene i and sample j and T'_{jk} be the score of the mutation occurring in position k and sample j. Then,

$$
\begin{aligned}
F_{ij}(x) &= P(T_{ij} < x | Y_{ij} = 1, M_0) = \frac{P(T_{ij} < x, Y_{ij} = 1 | M_0)}{P(Y_{ij} = 1 | M_0)} \\
&= \frac{\sum_{m \geq 1} P(T_{ij} < x, W_{ij} = m | M_0)}{\sum_{m \geq 1} P(W_{ij} = m | M_0)} \\
&\approx \frac{P(T_{ij} < x, W_{ij} = 1 | M_0)}{P(W_{ij} = 1 | M_0)} \qquad \text{(since } P(W_{ij} \geq 2 | M_0) \approx 0\text{)} \\
&= \frac{\sum_{k \in G_i} P(T'_{jk} < x, X_{jk} = \text{nsm}, X_{jl} \neq \text{nsm}, l \in G_i, l \neq k | M_0)}{\sum_{k \in G_i} P(X_{jk} = \text{nsm}, X_{jl} \neq \text{nsm}, l \in G_i, l \neq k | M_0)} \\
&\qquad \text{(where nsm = nts, ntv, iid, or fid)} \\
&= \frac{\sum_{k \in G_i} P(T'_{jk} < x | X_{jk} = \text{nsm}, M_0)\, q_j b_k \prod_{l \in G_i}^{l \neq k} (1 - q_j b_l)}{\sum_{k \in G_i} q_j b_k \prod_{l \in G_i}^{l \neq k} (1 - q_j b_l)} \\
&\qquad \text{(where } b_k = e_k \hat{p}_{t_k} + f_k \hat{p}_{v_k} + \hat{p}_7 + \hat{p}_8\text{)} \\
&= \frac{\sum_{k \in G_i} P(T'_{jk} < x | X_{jk} = \text{nsm}, M_0)\, \frac{q_j b_k}{1 - q_j b_k}}{\sum_{k \in G_i} \frac{q_j b_k}{1 - q_j b_k}} \\
&\approx \frac{\sum_{k \in G_i} P(T'_{jk} < x | X_{jk} = \text{nsm}, M_0)\, q_j b_k}{\sum_{k \in G_i} q_j b_k} \qquad \text{(since } 1 - q_j b_k \approx 1\text{)} \\
&= \frac{\sum_{k \in G_i} P(T'_{jk} < x | X_{jk} = \text{nsm}, M_0)\, b_k}{\sum_{k \in G_i} b_k}
\end{aligned}
$$

Results for the refined background mutation model

We separate the rates of mutations according to the mutation types (transition or transversion), base pair types ($A:T$ or $G:C$) and their context (CpG dinucleotide contexts). However, we did not separate the rates of the two types of mutation for each transversion: $A:T \rightarrow C:G$, $A:T \rightarrow T:A$ for the transversion at $A:T$ and $C:G \rightarrow A:T$, $C:G \rightarrow G:C$ for the transversion at $C:G$ in non CpG or in CpG. As discussed in section 4 of the main manuscript, we have modified the method

Table 1: Result for simulated data

Sample variation	Cutoff	Average number	Original method	New method	Ding's method
Moderate	0.005	TP	12.9	13.6	9.9
		FP	1.3	1.8	1.7
	0.01	TP	14.9	15.4	11.7
		FP	3	3.7	3.4
High	0.005	TP	13.4	14.1	9.9
		FP	0.2	0.4	2.0
	0.01	TP	15.1	15.7	11.7
		FP	0.6	1	3.9

TP : true positives, FP : false positives

so that each of them has a separate mutation rate. This increases the number of parameters by 3. We applied this new method as well as our original method and Ding's method to the simulated data generated described in section 3.1 of the main manuscript. For 200 repeated simulations, we calculated the average number of true and false positives for the three methods and presented them in Table 1. It shows that the new method increases true positives as well as false positives compared to the original method.

References

Ding, L., Getz, G., Wheeler, D. A., Mardis, E. R., Mclellan, M. D., Cibulskis, K., Sougnez, C., Greulich, H., Muzny, D. M., Morgan, M. B., Fulton, L., Fulton, R. S., Zhang, Q., Wendl, M. C., Lawrence, M. S., Larson, D. E., Chen, K., Dooling, D. J., Sabo, A., Hawes, A. C., Shen, H., Jhangiani, S. N., Lewis, L. R., Hall, O., Zhu, Y., Mathew, T., Ren, Y., Yao, J., Scherer, S. E., Clerc, K., Metcalf, G. A., Ng, B., Milosavljevic, A., Gonzalez-Garay, M. L., Osborne, J. R., Meyer, R., Shi, X., Tang, Y., Koboldt, D. C., Lin, L., Abbott, R., Miner, T. L., Pohl, C., Fewell, G., Haipek, C., Schmidt, H., Dunford-Shore, B. H., Kraja, A., Crosby, S. D., Sawyer, C. S., Vickery, T., Sander, S., Robinson, J., Winckler, W., Baldwin, J., Chirieac, L. R., Dutt, A., Fennell, T., Hanna, M., Johnson, B. E., Onofrio, R. C., Thomas, R. K., Tonon, G., Weir, B. A., Zhao, X., Ziaugra, L., Zody, M. C., Giordano, T., Orringer, M. B., Roth, J. A., Spitz, M. R., Wistuba, I. I., Ozenberger, B., Good, P. J., Chang, A. C., Beer, D. G., Watson, M. A., Ladanyi, M., Broderick, S., Yoshizawa, A., Travis, W. D., Pao, W., Province, M. A., Weinstock, G. M., Varmus, H. E., Gabriel, S. B., Lander, E. S., Gibbs, R. A., Meyerson, M., and Wilson, R. K. (2008). Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*, **455**(7216), 1069–1075.